Comparing electoral campaigns by analysing online data

Javier A. Espinosa-Oviedo[156], Genoveva Vargas-Solar[256], Vassil Alexandrov[134], Géraldine Castel[7]
[1] *BSC, Barcelona Supercomputing Centre, Spain*
[2] *CNRS, French Council of Scientific Research, France*
[3] *ICREA, Catalan Institution for Research and Advanced Studies, Spain*
[4] *ITESM, Tecnológico de Monterrey, Mexico*
[5] *LAFMIA, French-Mexican Laboratory of Informatics and Automatic Control, France*
[6] *LIG, Laboratory of Informatics of Grenoble, France*
[7] *Université Stendhal, Grenoble 3, France*
{espinosa, gvargas}@imag.fr, vassil.alexandrov@bsc.es, geraldine.castel@u-grenoble3.fr

Abstract- *The use of information and communication technologies (ICT) in the political sphere is nowadays a key aspect for running electoral campaigns. Thus, our work addresses the influence of ICT and candidate practices during electoral campaigns. Our approach is based in the collection of data produced by political candidates so that experts can analyse them through an analytics processes. Accordingly, this paper presents results concerning three of the data collections life cycle phases: collection, cleaning, and storage. The result is a data collection ready to be analysed for different purposes. The paper also describes our experimental validation for comparing political campaigns behaviour in France and the United Kingdom during the European elections in 2014.*

I. INTRODUCTION

The use of ICTs for political purposes is a relatively new research field as the first publications date from 1980s [1], [2]. This continuous evolution of tools has been paralleled by a shift in attention from sites to forums [3], [4], blogs [5], [6], or social networks [7], [8]. The problem of integrating different data sources for supporting the analytics processes is not new in the database domain [9]. Most proposals assume that data providers (heterogeneous or not) are known in advance [10] and thus integration is based on knowledge about the data structure [11], content, semantics [12] and constraints. However, the emergence of new kinds of data providers like services (e.g. Twitter, Facebook), where there are no schemas, introduced new challenges [11]. Data also started to acquire "new" properties (more volume, velocity, variety) and with them emerged the need of building huge curated data collections [13], [14]. The challenge is thus to collect political data continuously [15] in order to analyze the influence of ICT during electoral campaigns.

II. OVERVIEW OF THE APPROACH

Figure 1 shows the overview of our approach. This process is recurrently executed since new data is produced.

1.Data collection. Data is collected according to different modes (push, pull) and at different rates when data are produced continuously. In the case of Web pages and blogs we collect their content using crawling tools and Web scrapping techniques.

2.Data analytics. For each attribute of a given data structure we identify the distribution of the values within the collected data. We consider some level of uncertainty so we identify missing values and infers some proposals based on computed values distributions (e.g., using extrapolation), as well as discovering possible relations among data attributes (e.g., equivalence, functional dependency, temporal or casual correlations). This phase generates views that provide an abstract representation of the data collection contents.

3.View storage. Depending on the characteristics of the data, views can be materialized and stored together with raw data. These decisions consider the probability of data to be accessed and processed together based on their possible dependencies. Data organization can ensure performance and reduction of memory and communication resources consumption during the data analytics processes.

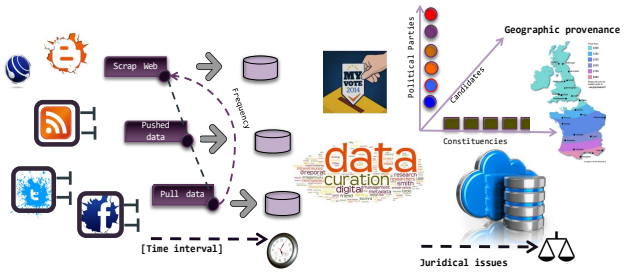

Figure 1 Data collection and curation overview

III. BUILDING AND MAINTAINING DATA VIEWS

The main idea of our approach is not to transform collected data but to generate an abstract aggregated view and then tag it with information that can be used for further data processing tasks. In this sense views can be seen as a materialized world, but extracted *a posteriori* after having created a database. As shown in the class diagram of figure 2, a View characterizes the content provided by a given dataProvider as a document composed of set of attributes, where an Attribute provides a snapshot of a given attribute's values domain for a given dataset. An attribute within the dataset has maximum and minimum values, a standard deviation of the values assigned to the attribute in the different documents collected in the dataset, and the variation of values across the dataset elements represented by a histogram. Within a dataset an attribute can

have null and missing values that must be inferred in order to characterize its domain type as precisely as possible. Indeed, many data collections represent missing values by dummy values and therefore we want to represent those cases.

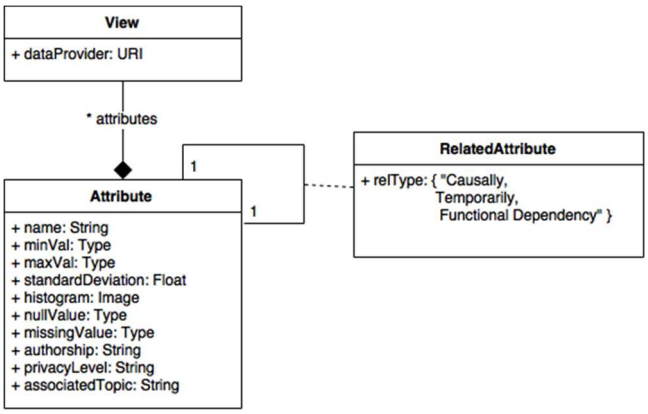

Figure 2 UML class diagram of the concept of View

IV. Validation

We built a system to analyse and compare campaigns in UK and France of the European elections in 2014 based on our approach (cf. We collected 30Gb of data comprising 12 parties and 100 candidates in France and UK, and they concern only online activities reported in Twitter, Facebook and official sites, pages and blogs. We used JSON as data model and we then implemented document processing tasks to characterize the content of collected data.

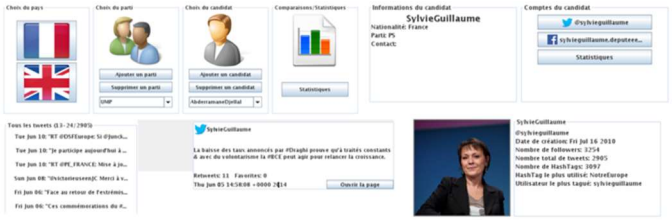

Figure 3 Profiling a candidate's campaign on social networks

V. Conclusions

This paper introduced our approach for building and curating political data collections and preparing them for the analytics process. Our first contribution in this paper regards the strategies used for characterizing and inferring data content through the notion of view. Some inference had to deal with uncertainty that we addressed associating accuracy probabilities to inferences so as to guide the data scientist in her further data analytics design.

References

[1] J. B. Abramson, F. C. Arterton, and G. R. Orren, "The Electronic Commonwealth: The Impact of New Media Technologies on Democratic Politics," *Michigan Law Review*, vol. 87, no. 6, p. 1393, May 1989.

[2] J. D. H. Downing, "Computers for Political Change: PeaceNet and Public Data Access," *Journal of Communication*, vol. 39, no. 3, pp. 154–162, Sep. 1989.

[3] S. Wojcik, "Les forums électroniques municipaux,espaces de débat démocratique ?," *Sciences de la Société: Démocratie locale et Internet*, no. 60, pp. 107–125, 2003.

[4] M. Marcoccia, "Les webforums des partis politiques français : quels modèles de discussion politique ?," *Mots Les langages du politique [En ligne], URL: http://motsrevuesorg/512*, pp. 49–60, 2006.

[5] F. Greffet, "Les blogues politiques. Enjeux et difficultés de recherche à partir de l'exemple français," *Communication Information médias théories pratiques*, vol. 25, no. 2, pp. 200–211, 2007.

[6] S. Gadras, "Public Sphere and Political Communication: how Does the Public Sphere Evolves with the Development of ICTs in French Local Politics?," in *The European Public Sphere: From critical thinking to responsible action*, Brussels, Peter Lang, 2012.

[7] N. Jackson and D. Lilleker, "Microblogging, Constituency Service and Impression Management: UK MPs and the Use of Twitter," *The Journal of Legislative Studies*, vol. 17, no. 1, pp. 86–105, Mar. 2011.

[8] M. Margaretten and I. Gaber, "The Crisis in Public Communication and the Pursuit of Authenticity: An Analysis of the Twitter Feeds of Scottish MPs 2008-2010," *Parliamentary Affairs*, vol. 67, no. 2, pp. 328–350, Apr. 2014.

[9] X. L. Dong and D. Srivastava, "Big data integration," in *2013 IEEE 29th International Conference on Data Engineering (ICDE)*, 2013, vol. 6, no. 11, pp. 1245–1248.

[10] X. L. Dong, L. Berti-Equille, Y. Hu, and D. Srivastava, "Global detection of complex copying relationships between sources," *Proceedings of the VLDB Endowment*, vol. 3, no. 1–2, pp. 1358–1369, Sep. 2010.

[11] V. Cuevas-Vicenttín, J. L. Zechinelli-Martini, and G. Vargas-Solar, "Andromeda: Building e-Science Data Integration Tools," in *Proc. of the 17th Int. DEXA Conference*, 2006, pp. 44–53.

[12] F. Osborne and E. Motta, "Klink-2: Integrating Multiple Web Sources to Generate Semantic Topic Networks," in *Proc. of the 14th Int. Semantic Web Conference (ISWC 2015)*, 2015, pp. 408–424.

[13] M. Adiba, J. C. Castrejón, J. A. Espinosa-Oviedo, G. Vargas-Solar, and J.-L. Zechinelli-Martini, "Big Data Management: Challenges, Approaches, Tools and their limitations," in *Networking for Big Data*, CRC Press, 2015.

[14] A. Labrinidis and H. V. Jagadish, "Challenges and opportunities with big data," *Proceedings of the VLDB Endowment*, vol. 5, no. 12, pp. 2032–2033, Aug. 2012.

[15] M. Ma, P. Wang, and C.-H. Chu, "Data Management for Internet of Things: Challenges, Approaches and Opportunities," in *2013 IEEE International Conference on Green Computing and Communications and IEEE Internet of Things and IEEE Cyber, Physical and Social Computing*, 2013, pp. 1144–1151.